Обучение машинного перевода без параллельных текстов*

Строганов А. А. Бахтеев*[1] *О. Ю. Стрижов*[2] *В. В.
[1]Московский физико-технический институт
[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Данная работа посвящена исследованию машинного перевода без использования параллельных текстов. Исследование сконцентрировано на использовании нейросети с несколькими моделями Seq2seq для перевода с одного языка на другой и обратно. Особенностью данных моделей является то, что исходный текст кодируется во внутреннее представление модели, а затем декодируется в текст на другом языке. Две модели Seq2seq имеют общее скрытое пространство. Примером, иллюстрирующим работоспособность данного алгортма, будет использован эксперимент по взаимному переводу с двух похожих языков – русского и украинского.

Ключевые слова: *машинный перевод, нейросеть, Seq2seq.*

1 Введение

Благодаря недавним достижениям в области глубокого обучения и наличию крупномасштабных параллельных корпусов, машинный перевод достиг впечатляющей производительности на нескольких языковых парах. Тем не менее, эти модели работают очень хорошо, только если они снабжены огромным количеством параллельных данных в порядке миллионов параллельных предложений. К сожалению, параллельные корпуса стоят дорого [2], поскольку они требуют специализированного опыта и часто не существуют для языков с низким уровнем ресурсов.

Есть несколько подходов к построению оптимального метода обучения. Предлагается использовать рекуррентные нейронные сети с короткой и долгой памятью и нейронные сети, в которых реализовано внимание. В других методах используются нейронные сети, которые осуществляют перевод в два этапа. Такой метода называется Seq2Seq [3].

Данная работа посвящена последнему методу последовательного перевода. Предлагается с помощью первой рекуррентной нейронной сети, основанной на долгой памяти перевести входящую последовательность в вектор, а с помощью второй перевести этот вектор в выходную последовательность на нужном нам языке [1]. Данный метод позволяет гораздо быстрее обучить нейронную сеть переводу с одного языка на другой, в связи с использованием ей предыдущего опыта и наличию у нее памяти и внимания. Проверка и анализ метода проводятся с помощью алгоритма BLEU(Bilingual evaluation understudy) для проверки качества текста, переведенного с одного языка на другой на паре языков русский-украинский.

2 Постановка задачи

Во время обучения нет параллельных пар предложений. Предполагаем, что нам подойдет модель, отображающая предложения из обоих языков в одно общеее векторное пространство.

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Авторы: А.В. Грабовой, О.Ю. Бахтеев, В.В. Стрижов, Eric Gaussier, координатор Малиновский Г.С. Консультант: Бахтеев О. Ю.

Мы будем использовать модель, в которой используются главных юнита: encoder f и decoder g. f и g в нашей конкретной модели - две рекуррентные нейронные сети. Задача f - отображать предложения в латентное пространство(сразу для двух языков) и g - отображать из латентного пространства в предложения(первого языка и второго соответственно). Введем обозначения: $D^{src} = [s_1^{src}, ..., s_{m_{src}}^{src}]$, $D^{tgt} = [s_1^{tgt}, ..., s_{m_{tgt}}^{tgt}]$.

Для реализации этого метода определим функционалы, которые будут минимизироваться. Чтобы модель не обучилась возвращать в конце цикла исходные данные, необходимо зашумить исходные предложения. Пусть $\sigma(x)$ - результат наложения шума на слово x. Оптимизировать будем следующую функцию:

$$L_{AE} = ||d(e(\sigma(x))) - x||^2$$

Пусть дана какая-то модель слабого перевода $\hat{g}$. На втором шаге функция потерь будет иметь вид:

$$L_{TR} = ||d(e(\hat{g}(e(x))) - x||^2$$

Пусть дана модель D, различающая скрытые представления векторов предложений из двух языков. На последнем шаге оптимизируем дискримантор, чтобы он различал представления векторов разных языков в скрытом пространстве:

$$L_{ADV} = \log p(lang = src|Encoder(x)) + \log p(lang = tgt|Encoder(y))$$

В итоге нужно минимизировать следующую функцию:

$$L = a * L_{AE} + b * L_{TR} + c * LADV \longrightarrow min$$

здесь a, b, c - калибруемые гиперпараметры.

3 Постановка задачи

Даны обучающая выборка $\mathfrak{D}^{\text{src}} = \{\mathbf{s}_1^{\text{src}}, \ldots, \mathbf{s}_{m_{\text{src}}}^{\text{src}}\}$, $\mathfrak{D}^{\text{tgt}} = \{\mathbf{s}_1^{\text{tgt}}, \ldots, \mathbf{s}_{m_{\text{tgt}}}^{\text{tgt}}\}$, которая состоит из двух корпусов произвольных предложений для каждого языка; и валидационная выборка $\mathfrak{D}^{valid} = \{(\tilde{\mathbf{s}}_1^{\text{src}}, \tilde{\mathbf{s}}_1^{\text{tgt}}), \ldots, (\tilde{\mathbf{s}}_{m_{\text{valid}}}^{\text{src}}, \tilde{\mathbf{s}}_{m_{\text{valid}}}^{\text{tgt}})\}$, которая состоит из пар параллельных предложений.

Задача состоит в том, чтобы, используя обучающую выборку, построить модель M перевода предложений с языка src на язык tgt и выбрать ее параметры таким образом, чтобы максимизировать среднее значение метрики BLEU на валидационной выборке

$$\frac{1}{m_{\text{valid}}} \sum_{i=1}^{m_{\text{valid}}} \text{BLEU}(M(\tilde{\mathbf{s}}_i^{\text{src}}), \tilde{\mathbf{s}}_i^{\text{tgt}}) \to \max_M$$

3.1 Гипотеза

Пусть $\mathcal{S}^l$ – множество всех предложений на языке $l \in \{\text{src}, \text{tgt}\}$. Выдвигается гипотеза о том, что существует единое латентное пространство $\mathcal{L}$ и отображения $\boldsymbol{f}^l : \mathcal{S}^l \to \mathcal{L}$ и $\boldsymbol{g}^l : \mathcal{L} \to \mathcal{S}^l$ такие, что

- для $\forall l_1, l_2 \in \{\text{src}, \text{tgt}\}$ и $\forall s^{l_1} \in \mathcal{S}^{l_1}$ $\boldsymbol{g}^{l_2}(\boldsymbol{f}^{l_1}(s^{l_1}))$ совпадает с s^{l_1}, если l_1 совпадает с l_2, и является корректным переводом s^{l_1}, если l_1 и l_2 различаются
- распределения образов $\boldsymbol{f}^{\text{src}}(\mathcal{S}^{\text{src}})$ и $\boldsymbol{f}^{\text{tgt}}(\mathcal{S}^{\text{tgt}})$ совпадают

3.2 Описание метода

Предлагаемый метод заключается в том, чтобы для каждого $l \in \{\text{src}, \text{tgt}\}$ моделировать отображения $\boldsymbol{f}^l$ и $\boldsymbol{g}^l$ кодировщиком $\mathbf{f}^l$ и декодировщиком $\mathbf{g}^l$ соответственно. Таким образом, моделью перевода является композиция $\mathbf{g}^{\text{tgt}} \circ \mathbf{f}^{\text{src}}$.

Оптимизация проводится следующим образом. Функция ошибки содержит три слагаемых, которые соответствуют сделанным предположениям об отображениях $\boldsymbol{f}^l$ и $\boldsymbol{g}^l$.

3.3 Ошибка восстановления

Для каждого $l \in \{\text{src}, \text{tgt}\}$ рассматривается входное предложение $\mathbf{s}^l$ на языке l. Ошибки считаются между $\mathbf{s}^l$ и его образом при отображении $\mathbf{g}^l \circ \mathbf{f}^l$.

3.4 Ошибка перевода

Рассматривается предложение $\mathbf{s}^{\text{src}}$ без ограничения общности на языке src. В качестве входного предложения используется $\mathbf{s}^{\text{tgt}}$ – перевод $\mathbf{s}^{\text{src}}$, полученный с помощью некоторой слабой модели перевода M^0. Ошибка перевода с языка tgt на язык src считается между $\mathbf{s}^{\text{src}}$ и образом $\mathbf{s}^{\text{tgt}}$ при отображении $\mathbf{g}^{\text{src}} \circ \mathbf{f}^{\text{tgt}}$. Аналогичным образом считается ошибка перевода с языка src на язык tgt. В качестве M^0 используется пословный переводчик на основе предобученных векторных представлений слов $\mathcal{E}^{\text{src}} = \{\mathbf{x}_i\}_{i=1}^{n_{\text{src}}}$ и $\mathcal{E}^{tgt} = \{\mathbf{y}_i\}_{i=1}^{n_{\text{tgt}}}$. Переводом слова $\mathbf{x} \in \mathcal{E}^{\text{src}}$ является $\mathbf{y} = \arg\min_{\mathbf{y} \in \mathcal{E}^{\text{tgt}}} \rho(\mathbf{x}, \mathbf{y})$, где ρ – косинусное расстояние (перевод слов языка tgt осуществляется аналогично).

3.5 Штраф за различие распределений

Вводится дискриминатор $\mathbf{d}$, который решает задачу классификации векторов $\mathbf{h}^l$ латентного пространства $\mathcal{L}$ на классы $\{0, 1\}$: $\mathbf{h}^l \in 0 \Leftrightarrow l = \text{src}$. Векторы $\mathbf{h}^l$ получаются кодировщиком $\mathbf{f}^l$, в качестве промежуточного результата при отображениях входных предложений. В функцию ошибки добавляется штраф за то, что дискриминатор точно определяет язык входного предложения. Таким образом параметры модели оптимизируются таким образом, чтобы усложнить задачу дискриминатору. В свою очередь параметры дискриминатора оптимизируются параллельно параметрам модели. Соревновательный процесс оптимизации мотивирован желанием добиться сходства распределений латентных векторов для разных языков.

3.6 Детали метода

Введем словарь V^{src}, содержащий проиндексированные слова, встречающиеся в предложениях из $\mathfrak{D}^{\text{src}}$ и переводах предложений из $\mathfrak{D}^{\text{tgt}}$, получаемых с помощью M^0. Аналогично введем V^{tgt}.

Предложения на разных языках, использующиеся при оптимизации кодируются с помощью соответствующих словарей. Входные предложения зашумляются преобразованием σ: сначала с некоторой вероятностью q из них удаляется каждое слово, а затем производится случайная перестановка оставшихся слов с условием, что слово не может оказаться дальше чем на k позиций от своей начальной позиции (q, k – гиперпараметры).

Кодировщик $\mathbf{f}^l$ включает в себя слой векторного представления слов $\mathbf{e}^l$, параметры которого инициализируются векторами $\mathcal{E}^l$, и реккурентную нейронную сеть $\mathbf{r}^{\text{enc}}$. Декодировщик $\mathbf{g}^l$ включает в себя $\mathbf{e}^l$, $\mathbf{r}^{\text{dec}}$ и классификатор $\mathbf{c}^l$, который решает задачу многоклассовой классификации выходов $\mathbf{r}^{\text{dec}}$ на классы, соотвествующие словам в словаре V^l (отображает выходы $\mathbf{r}^{\text{dec}}$ в векторы вероятностей размерности $|V^l|$). Параметры $\mathbf{r}^{\text{enc}}$ и $\mathbf{r}^{\text{dec}}$ общие для $\mathbf{f}^{\text{src}}$, $\mathbf{f}^{\text{tgt}}$ и $\mathbf{g}^{\text{src}}$, $\mathbf{g}^{\text{tgt}}$ соответственно.

В качестве меры ошибок классификации используется кросс-энтропия CE. Итоговый вид функции ошибки

$$L_{\text{tran}} = w_1 \cdot \sum_{l \in \{src,tgt\}} CE(\mathbf{g}^l(\mathbf{f}^l(\sigma(\mathbf{s}^l))), \mathbf{s}^l) +$$

$$+ w_2 \cdot \sum_{l_1 \neq l_2 \in \{src,tgt\}} CE(\mathbf{g}^{l_2}(\mathbf{f}^{l_1}(\sigma(\mathbf{s}^{l_1}))), \mathbf{s}^{l2}) + w_3 \cdot \sum_{l \in \{src,tgt\}} CE(\mathbf{d}(\mathbf{f}^l(\sigma(\mathbf{s}^l))), \mathbb{I}[l = \text{src}])$$

$$L_{\text{disc}} = \sum_{l \in \{src,tgt\}} CE(\mathbf{d}(\mathbf{f}^l(\sigma(\mathbf{s}^l))), \mathbb{I}[l = \text{tgt}]).$$

Литература

[1] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[2] Philipp Koehn. *Statistical machine translation.* Cambridge University Press, 2009.

[3] Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*, 2017.

Поступила в редакцию